Probabilistic prosodification of lexical versus grammatical words

Stephanie S Shih
shih@ucmerced.edu

Cognitive & Information Sciences
University of California, Merced

89th LSA | Portland, OR | 11 January 2015

Lexical versus grammatical words

- Phonological basis to lexical (content) versus grammatical (function) word division = stress reducibility.

- Basic observation:
  - Content words do not reduce.
  - Function words reduce.

- Surface differences reflect deeper differences:
  - in stress encoding (Selkirk 1984, 1996; Kaisse 1985; Inkelas & Zec 1993; a.o.)
  - in lexical accessbility (Segalowitz & Lane 2000; Bell et al. 2009; a.o.)

Lexical versus grammatical words

- Goal of the current talk
  - Examine the empirical basis of these assumed reducibility and word category differences:
  - Starting from the surface phonetic evidence, what are the emergent word categories?

Minority view on word categories

- Patterns of reduction reflect more fine-grained lexical divisions.
- e.g., Altenberg 1987
  - examined how often a word in a corpus of prepared monologue speech appears in a prosodically unmarked form, carrying no stress or intonation (i.e., pitch, amplitude).
  - Traditional set of content words behaved more homogeneously in stress behavior: less zero-accent potential in content words.
  - Function words behaved very heterogeneously: some function words are more like content words, other behaved more prototypically “function”-like in not carrying accent.

Altenberg 1987's

Scale of word classes

Reducibility	Part of speech
0–10%	adjectives, lexical verbs, common nouns, *wh*-adverbs, ordinals, quantifying pronouns
11–20%	particles, *well*, proper nouns, demonstrative pronouns, cardinals, predeterminers
21–30%	closed-class adverbs, other subordinators, *do*
31–40%	*have*
41–50%	relative pronouns, modal auxiliaries, quantifying postdeterminers, demonstrative determiners (pl)
51–60%	prepositions, *and*, correlative subordinators
61–70%	*that*, *but*, demonstrative determiners (sg)
71–80%	*be*, personal pronouns
81–90%	possessive determiners, existential *there*
91–100%	infinitive marker, articles

Minority view on word categories

- Hirschberg (1993)
  - evidence from pitch accent
  - finer-grained word categories (4 categories) produced 9% accuracy improvement in predicting pitch accent, as compared to binary content versus function word grouping.

A methodological issue

- Exhaustive classification is crucial for research that requires prosodification information:
  - e.g., heavy-last or weight-driven syntactic alternations measured by stress (e.g., Anttila et al. 2010; Shih 2014)
  - e.g., measuring rhythmicity in a language (e.g., Temperley 2009)

This talk

- Examine patterns of reduction in monosyllabic English words using naturalistic speech data.
- Develop a less biased methodology for positing word categories based on reduction patterns.
- Results: Reduction-based classification of words points to 4-part division of lexicon,
  - Reduction-based classification of words → 4-part division of lexicon.
  - Factors (e.g., phrase structure, predictability) known to be sensitive to word categories are signficantly sensitive to these four classes.

Data

- Buckeye Variation in Conversation Corpus (Pitt et al.2007)
  - Phonetically-transcribed collection of conversational American English
  - ~ 1 hour conversations with 40 speakers in OH

Data

- Penn Treebank part of speech tags
  - by automatic parsing from Buckeye (Santorini 1990; Marcus et al. 1993)
  - random subset of monosyllabic words checked by hand and corrected.
- Duration and phonetic transcription from Buckeye

Data

- Dictionary pronunciation
  - coded from American English Unisyn lexicon (Fitt 2001)
  - + manual coding
- Corpus-internal word frequencies
- Phrase boundaries
  - approximated at silence and turn-taking tags, and preceding conjunctions *and*, *but*, and *if* (following Yao 2011; Anttila 2012)
- Previous and following conditional bigram probabilities
- Speech rate
  - = avg σ per second in phrase with target word (see e.g., Frank & Jaeger 2008)
- Accent ratio
  - how likely word is to carry pitch accent, taken from training corpus by Nenkova et al. 2007.

Data

- Disfluencies, interjections were removed.
- Data trimmed by removing all data in which word duration and average speech rate were more than three standard deviations from the log mean (following Kuperman & Bresnan 2012)
- *n* = 206,858 monosyllabic words

Quantifying reduction

- Two measures of reduction:

- Segmental distance
- Duration distance

Quantifying reduction:

Segmental distance

- A phonetic cue to stress reduction is deviation in segmental material.
  - e.g., full vowels → schwa
  - e.g., massive reduction in connected speech (Johnson 2004a): *divinity* [də.ˈvɪ.nə.ti] → [də.ˈvɪ̃.ə̪.ti]

Quantifying reduction:

Segmental distance

- Measure of segmental distance from Johnson (2004b):
  - = amount of phonetic deviation between citation form and transcribed pronunciation in the Buckeye corpus
  - based on a confusion matrix of perceptual similarity between phones.
  - normalized on scale from 0 to 2 (see Johnson 2004b).

Quantifying reduction:

Segmental distance

0 — 2

no reduction — maximal reduction

e.g.,	*things*	[θɪŋz]	= 0
		[θɪŋs]	= 0.123
		[θɛŋz]	= 0.146
		[θĩ.z]	= 0.44
		[sɪŋ.]	= 0.563
		[sĩ.ʒ]	= 0.911
		[h.ŋs]	= 1.05

Quantifying reduction:

Duration distance

- Another phonetic cue to stress reduction is duration (Fry 1955; Beckman & Edwards 1994; a.o.).
- Duration distance
  - = ratio of actual spoken duration and estimated citation durations of a given word, normalized by number of phones in both transcribed and citation pronunciations.
  - Estimated citation durations were taken by summing, for each phone in the dictionary-listed pronunciation of the word, the average duration of that phone across its every appearance in Buckeye.
  - (see Gahl et al. 2012 for similar methodology)

Quantifying reduction:

Duration distance

-
+
less reduction
more reduction

e.g.,	*can (Noun)*	= -0.00165	*a can of beans*
	can (Modal)	= 0.0235	*I can afford to…*

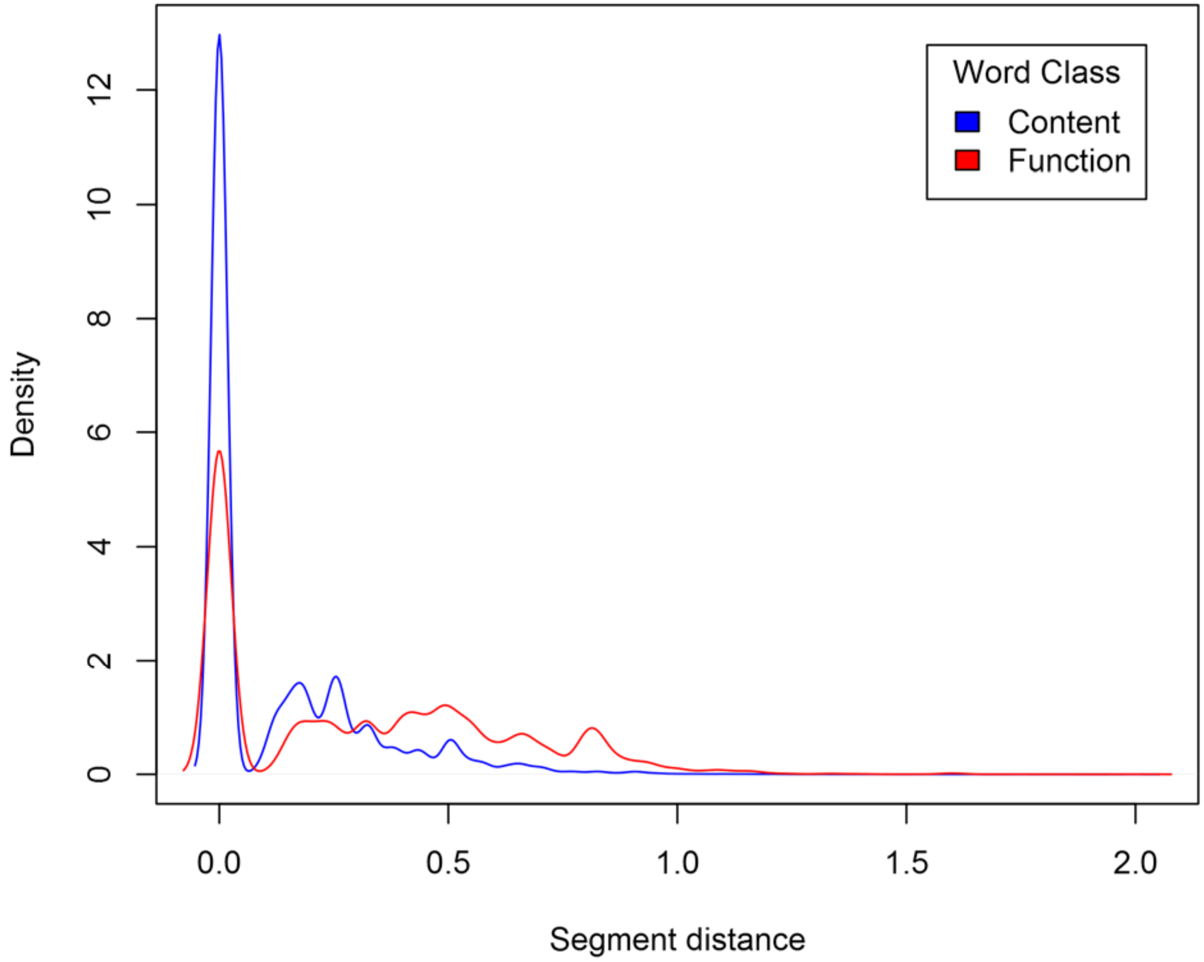
Preview of methodology

- Desideratum: a (relatively) unbiased way to discover viable hypotheses of word classifications that can be tested to see if they are meaningful lexical splits along content versus grammatical expectations.
- Word categories are likely to be meaningful if they
  - improve the power of analysis of reduction patterns without overfitting.
  - demonstrate independent differences as groups with respect to factors that should be sensitive to content and function word classes (e.g., structural position, predictability).
  - concord with theoretical expectations and/or previous findings.

Preview of methodology

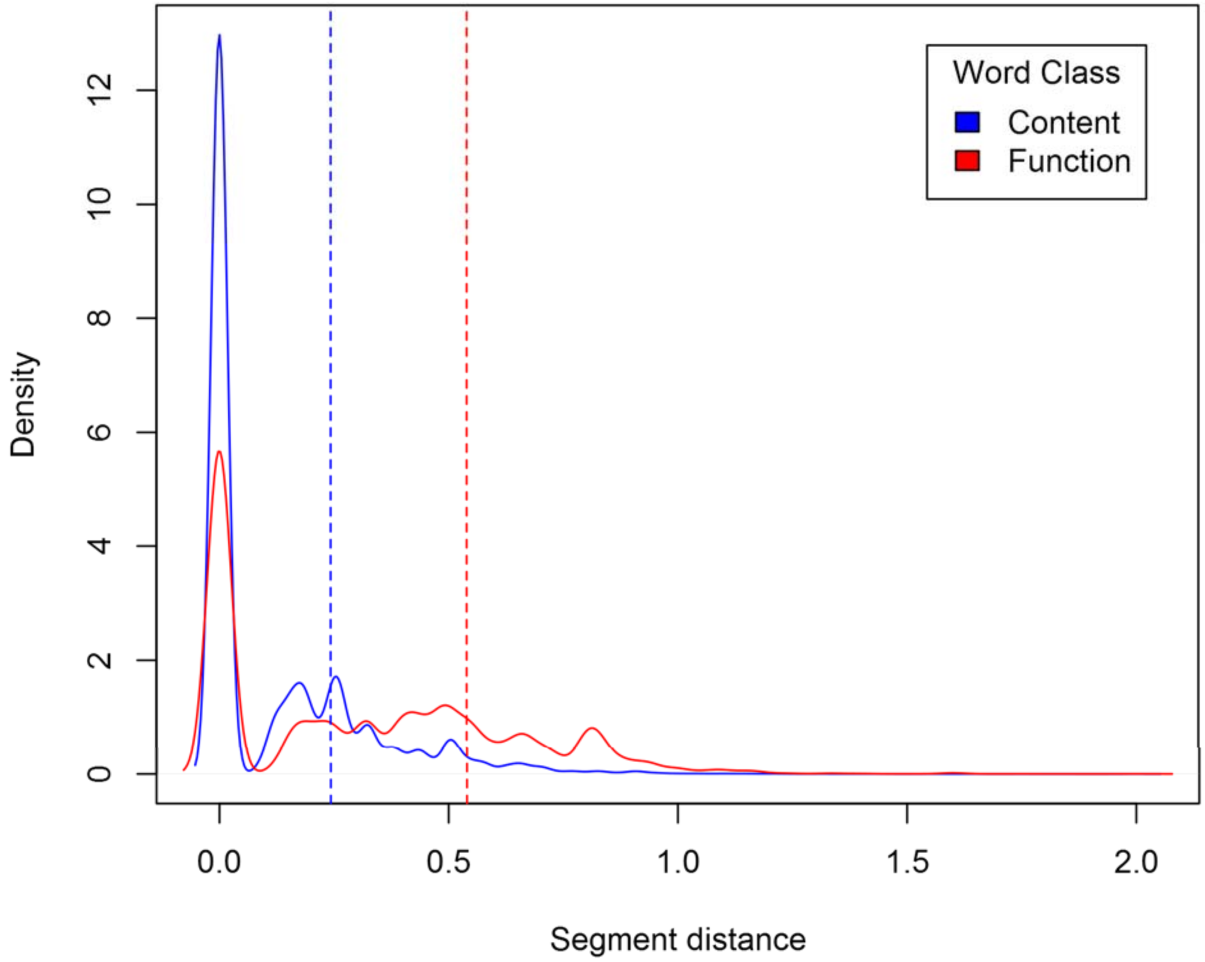
- Generate unbiased hypotheses of lexical/grammatical word categorizations based on observation of surface reduction patterns.
  - Hierarchical cluster analysis
- Test and compare hypotheses for their efficacy in lexical/grammatical distinctions.
  - Information-theoretic model comparison
  - Mixed effects modeling

Q_3 values



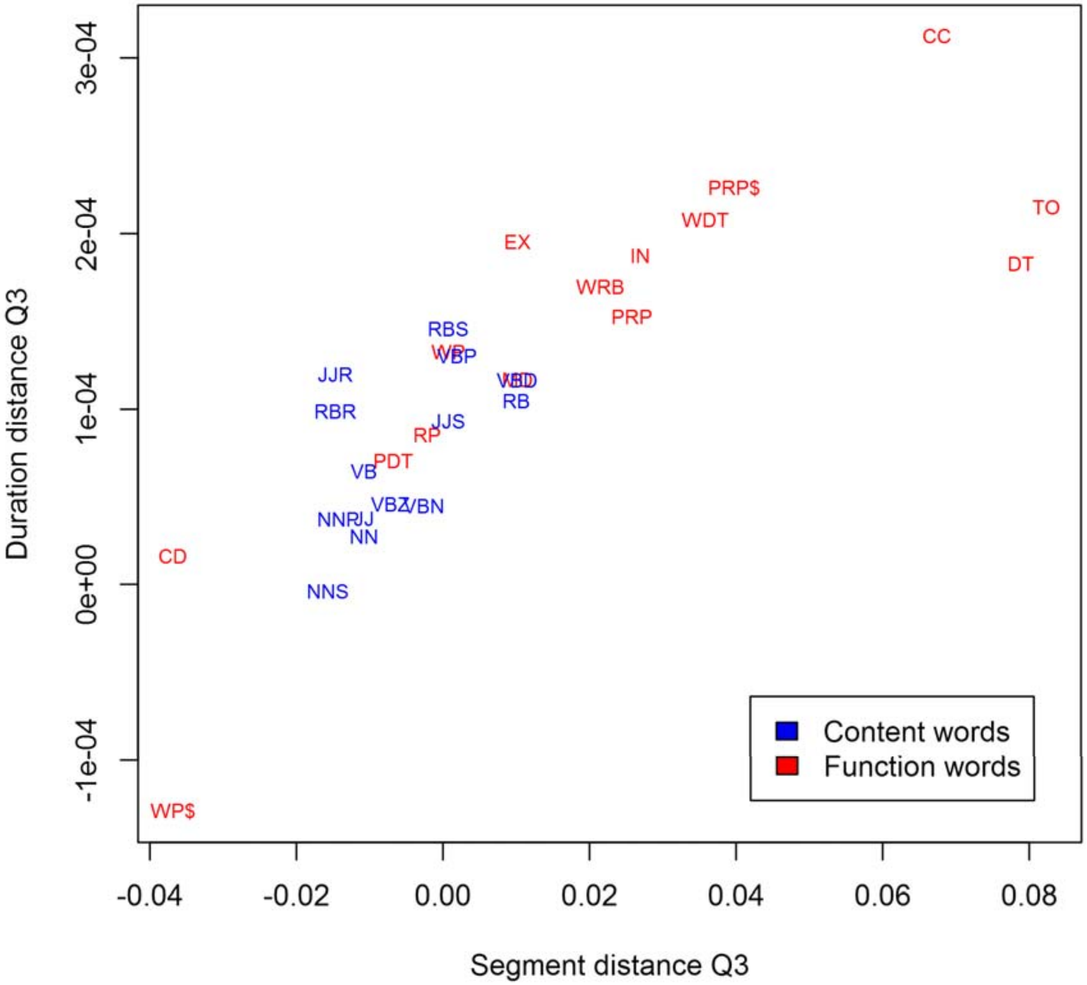
- Classes of words that exhibit more reduction should have:
  - Greater range of values in segment and duration distance measures.
  - More data mass located at higher values of these measures.

Q_3 values



- Captured by third quartile values (Q_3):
  - 75% of data occurs below Q_3.
  - 25% of data occurs above Q_3.

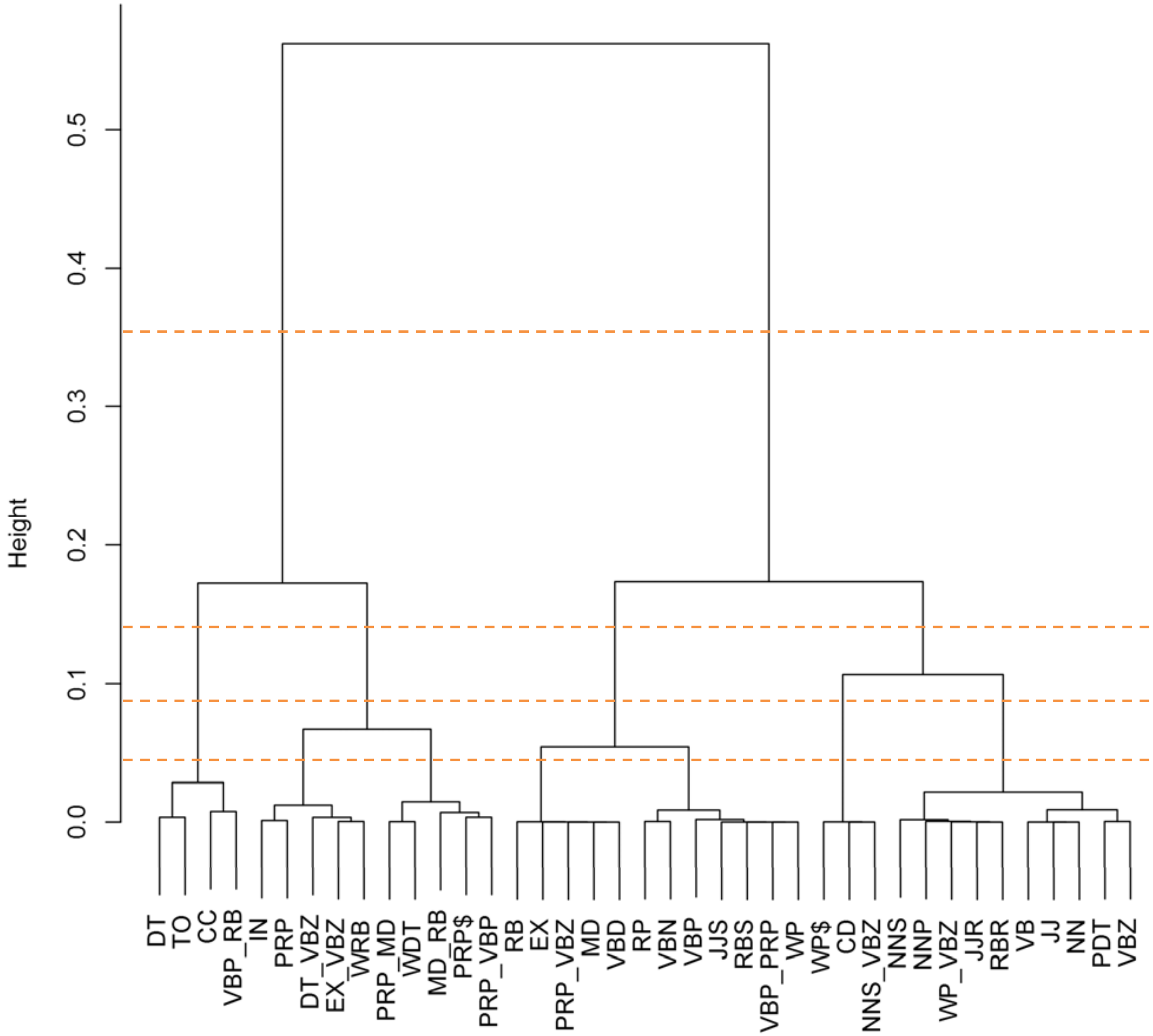
Q_3 values


CC
PRP$
TO
WDT
EX
IN
DT
WRB
PRP
RBS
WP
VBP
JJR
VBD
RB
RBR
JJS
RP
PDT
VB
VBZ
VBN
NNP
JJ
NN
CD
NNS
WP$
Content words
Function words
3e-04
2e-04
1e-04
0e+00
-1e-04
Duration distance Q3
-0.04
-0.02
0.00
0.02
0.04
0.06
0.08
Segment distance Q3

Hierarchical cluster analysis

- Bottom-up method of discovering groupings using Ward's method (Ward 1963):
  - pairs of clusters are formed by minimizing the increase of total within-cluster variance, as measured by sum-of-squares.
- Based on reduction patterns (segmental distance and duration distance) between groups.
- Provides numerous word category hypotheses.

Hierarchical cluster analysis



2 clusters

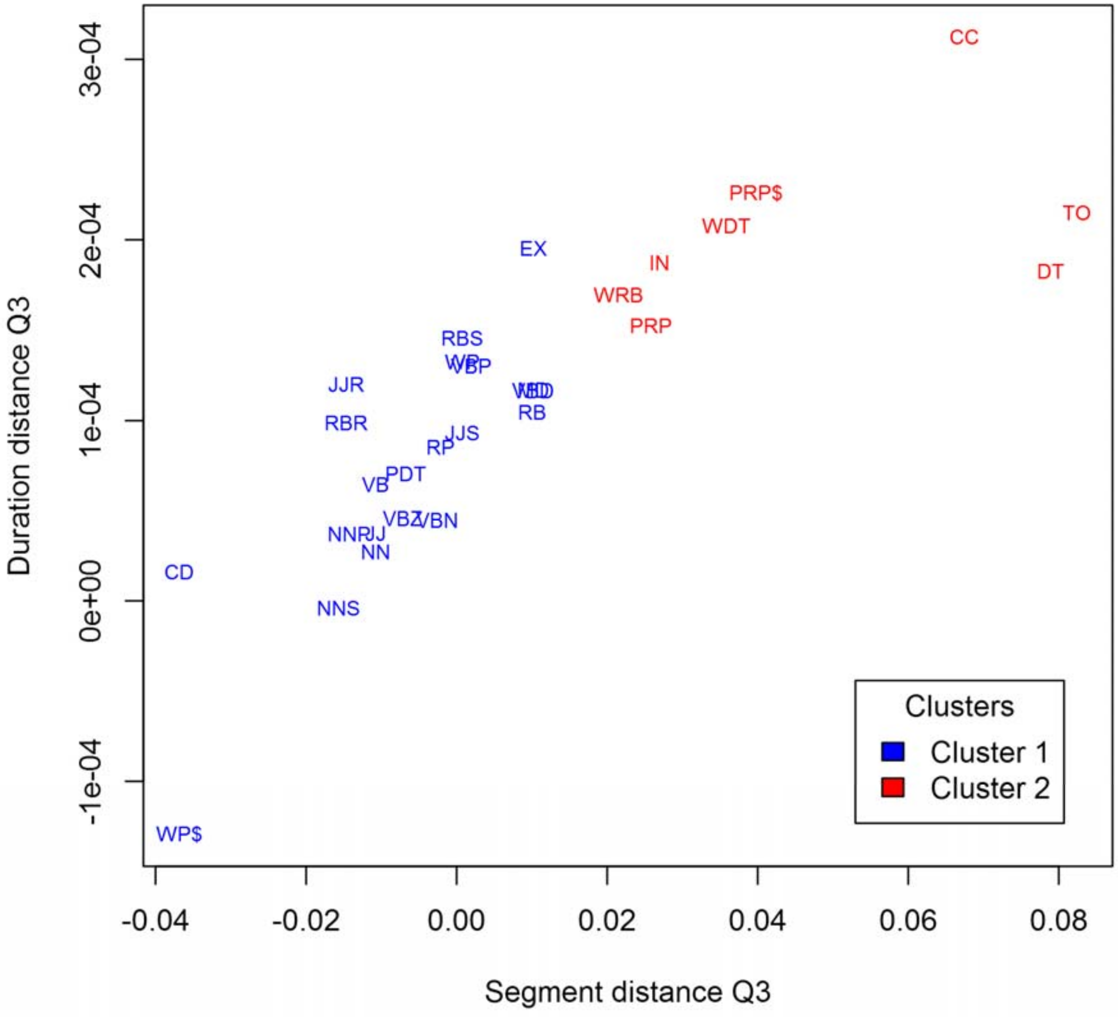
4 clusters

5 clusters

7 clusters

Hierarchical cluster analysis results:

2-cluster hypothesis

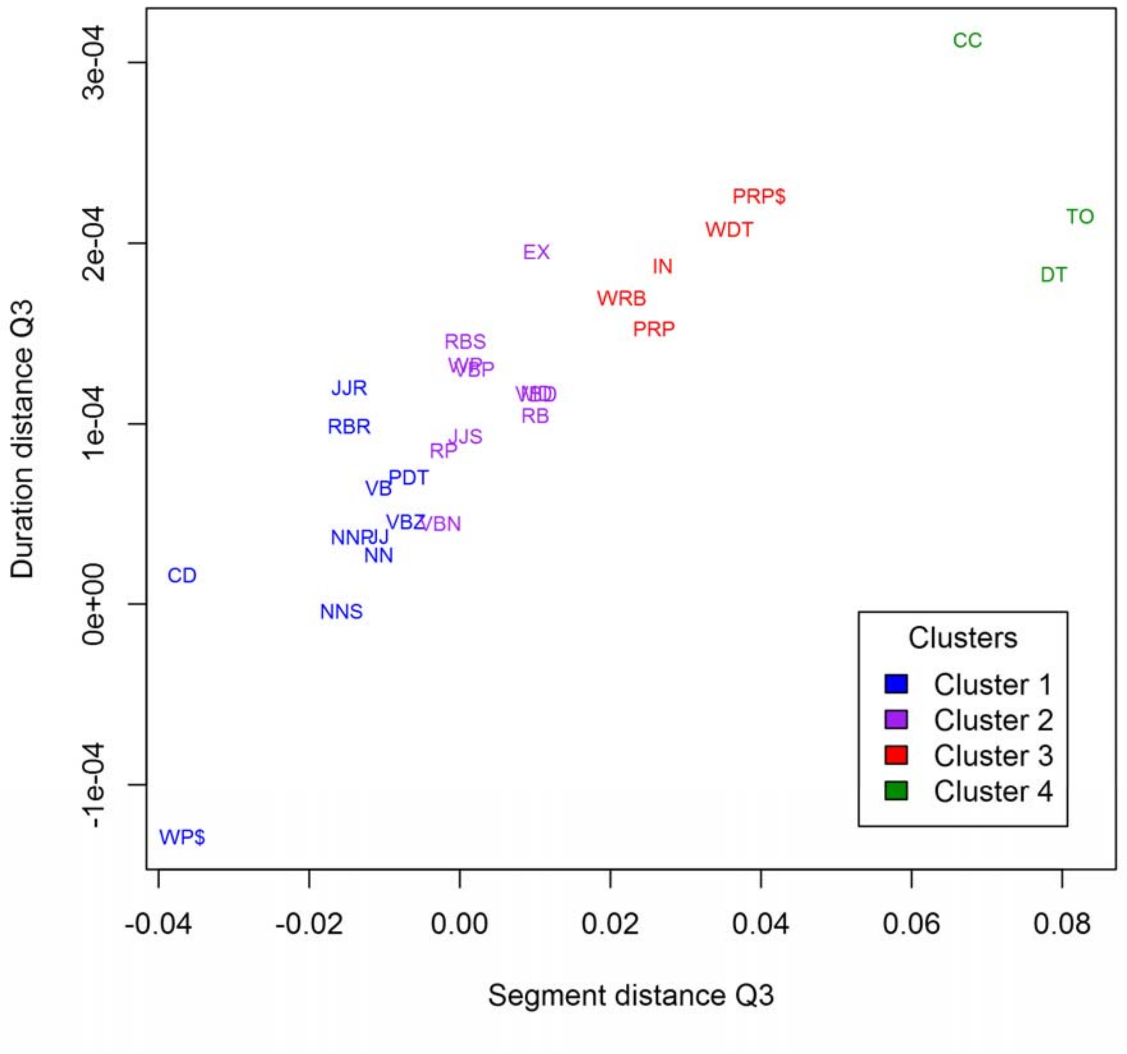


Cluster 1: cardinal numbers, possessive *wh* pronouns, adjectives, nouns, proper nouns, predeterminers, verbs, some adverbs, modal auxiliaries, adverbs, particles, *wh* pronoun, ex. *there*

Cluster 2: prepositions, pronoun, *wh* determiner, *wh* adverb, conjunctions, determiners, preposition *to*

Hierarchical cluster analysis results:

4-cluster hypothesis



Cluster 1: cardinal numbers, possessive *wh* pronouns, adjectives, nouns, proper nouns, predeterminers, some adverbs, present tense verbs (3p)

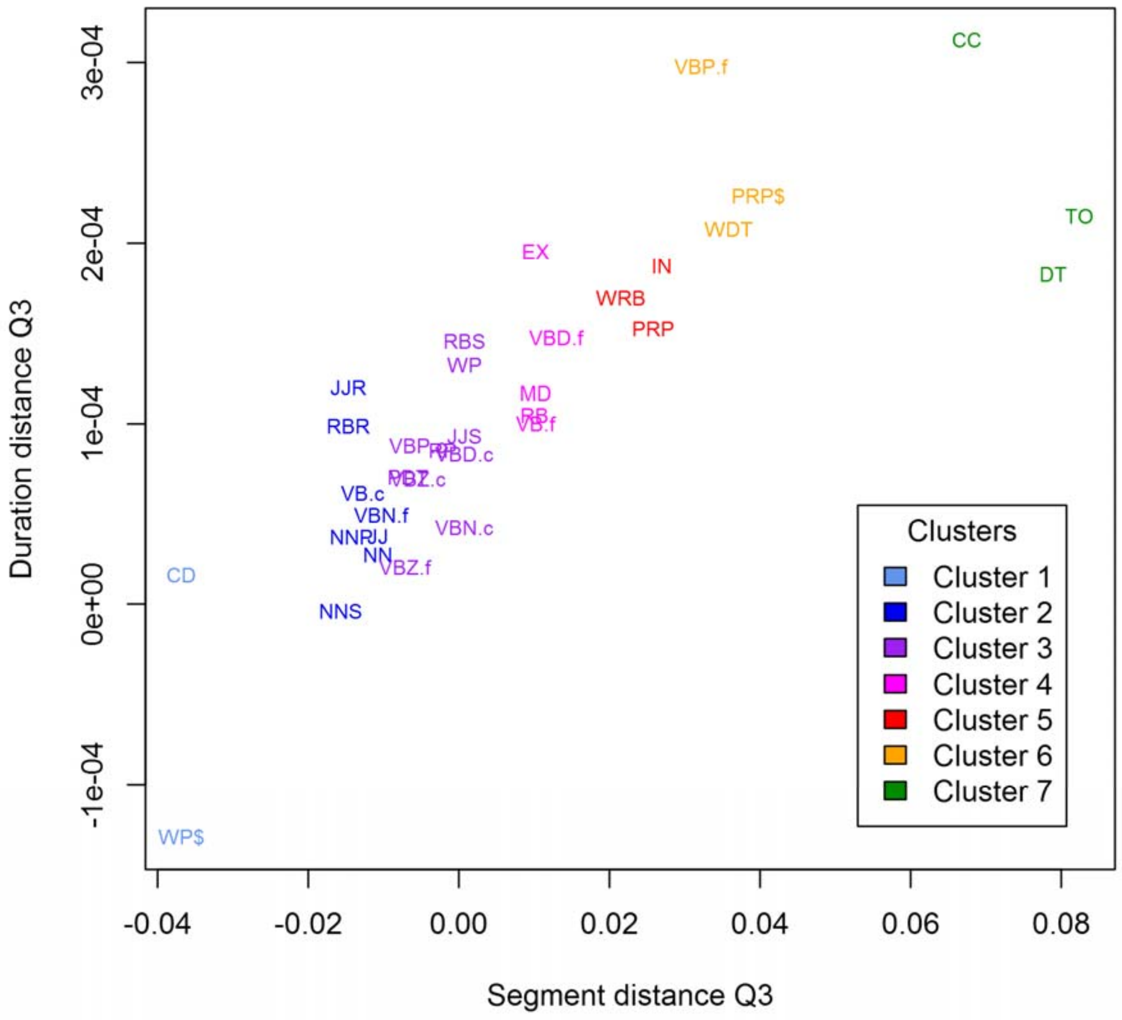
Cluster 2: modal auxiliaries, adverbs, particles, past tense verbs, present tense verbs (not 3p), *wh* pronoun

Cluster 3: preposition, pronoun, *wh* determiner, *wh* adverb

Cluster 4: conjunctions, determiners, preposition *to*

Hierarchical cluster analysis results:

7-cluster hypothesis



Cluster 1: cardinal numbers, possessive *wh* pronoun

Cluster 2: adjectives, nouns, proper nouns, present lexical verbs (3p), past participles *done, had, been*

Cluster 3: predeterminers, superlatives, particles, present tense lexical verbs, past tense lexical verbs present *is, has, does*, *wh* pronoun

Cluster 4: modal auxiliaries, ex. *there*, adverbs, past tense *had, were, was, did,* particles

Cluster 5: preposition, pronouns, *wh* adverb

Cluster 6: possessive pronouns, *wh* determiner, present verbs *are, have, do, am*

Cluster 7: conjunctions, determiners, preposition *to*

Testing categorization hypotheses

- Thirteen word cluster hypotheses:
  - Nine hypotheses from cluster modeling based on segmental and duration reduction
  - Standard content vs. function division
  - 10-part scale from Altenberg (1987)
  - Hybrid model with standard division + function word clusters
  - Null control model

Testing categorization hypotheses

- Tested in generalized linear mixed-effects models
  - Predicting reduction
  - Controls: speech rate, frequency, conditional bigram probabilities, accent ratio, prosodic/syntactic structure, phonological structure, word (random effect)

Testing categorization hypotheses

- Thirteen hypotheses compared using Akaike Information Criteria (AIC$_c$) model comparison (Burnham & Anderson 2002, 2004).

$$\mathrm{AIC}_c = -2\log\left(\mathcal{L}(\hat{\theta}|data)\right) + 2K$$

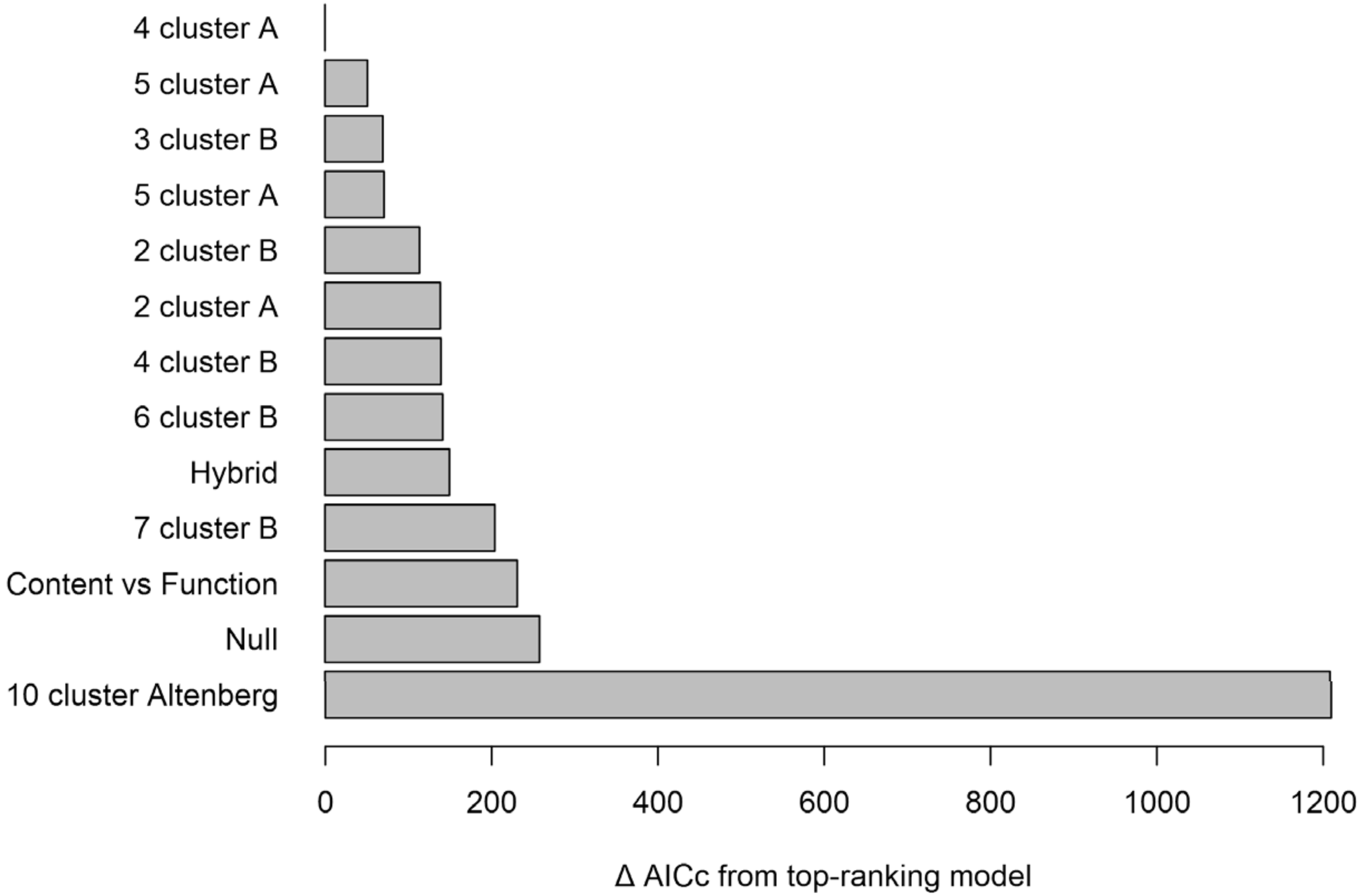
where $\left(\mathcal{L}(\hat{\theta}|data)\right)$ = likelihood of observed data given parameters $\hat{\theta}$,

and K = number of estimable parameters in the model.

Testing categorization hypotheses

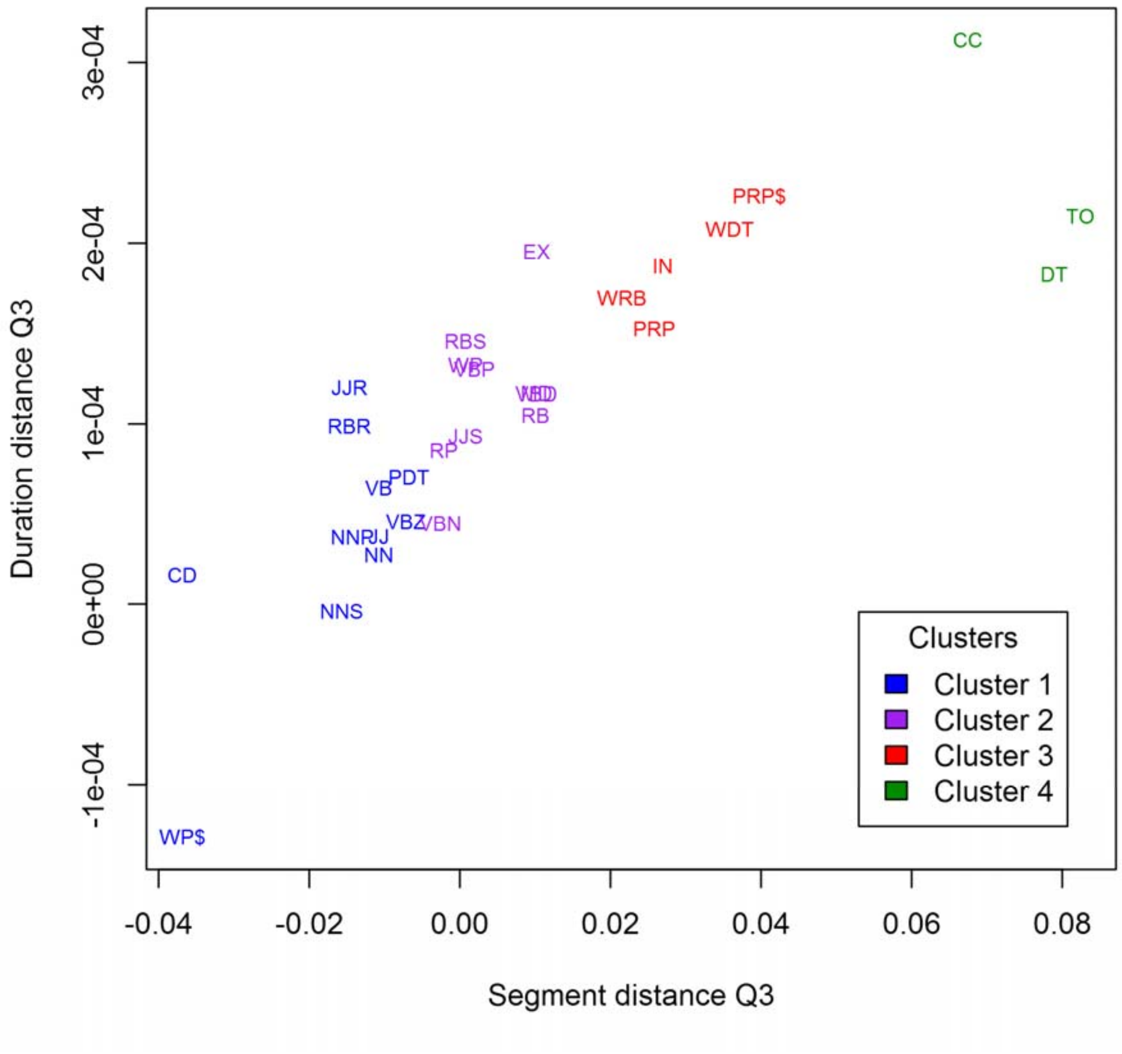
- AIC$_c$ model comparison measures how much weight of evidence there is for each hypothesis model.
- AIC$_c$ adjusts for number of parameters in a model, therefore does not necessitate nested models for comparison (vs. anovas, R^2, log-likelihood).
- Ranks models in order of which is closest to capturing full reality given the evidence available.

AIC$_c$ results on segmental distance



Hierarchical cluster analysis results:

4-cluster hypothesis



Cluster 1: cardinal numbers, possessive *wh* pronouns, adjectives, nouns, proper nouns, predeterminers, some adverbs, present tense verbs (3p)

Cluster 2: modal auxiliaries, adverbs, particles, past tense verbs, present tense verbs (not 3p), *wh* pronoun

Cluster 3: preposition, pronoun, *wh* determiner, *wh* adverb

Cluster 4: conjunctions, determiners, preposition *to*

4 categories of lexical/function words

- Are factors that are hypothesized to be sensitive to content/function word divisions also sensitive to divisions of these four clusters?
  - Frequency: content words are more sensitive to frequency-induced reduction (Bradley 1978; Bell et al. 2009; cf. Gordon & Caramazza 1982; Segalowitz & Lane 2000; a.o.)
  - Predictability: function words are predicted to be more sensitive to predictability-induced reduction than content words (Bell et al. 2009)
  - Phrase position: function words are expected to be more susceptible to phrase-internal reduction effects than content words (Selkirk 1984, 1996; Inkelas & Zec 1993; a.o.)

4 categories of lexical/function words

- Frequency
  - Cluster 2 words demonstrate significantly less of an effect for frequency-induced reduction than Cluster 1 words (β=-0.0002953, t=-2.197).
  - Cluster 4 words exhibit significantly less reduction when compared to other three clusters lumped together (β=-0.001083, t=-4.679).

4 categories of lexical/function words

- Predictability
  - Significant predictability differences for Clusters
  - For both preceding and following bigram predictability:
    - Particularly magnified for Cluster 4 words (β=0.0004068, t=-9.789).
  - For preceding bigram predictability:
    - Reduction increases across all clusters as probability of target word given preceding word increases.

4 categories of lexical/function words

- Phrase position (final vs. non-final)
  - Differences located primarily in Cluster 2 and 3 words.
  - Cluster 2 words demonstrate more phrase-internal-induced reduction than Cluster 1 words (β=0.002057, t=6.919).
  - Cluster 3 even more so (β=0.001348, t=7.572).

Conclusion

- Based on reduction patterns, how can lexical and grammatical word categories be determined, without *a priori* assuming such categories?
- Generated hypotheses using observations of surface patterns in natural speech.
- Tested hypotheses via model comparison.
- Examined validity of optimal hypothesis using comparison with expected effects across lexical divisions.

Conclusion

- Optimal four-way categorization presented largely mirrors expectations of content and function word categories, but also allows for more subtlety in dividing up the lexicon.
- Aligns with expectations of how content-like and function-like words might behave.
  - e.g., grammatical words that reduce less often and pattern more like lexical words are expected to are the ones that appear to carry more semantic information or are more closely tied to verb complex (e.g., auxiliaries).

Conclusion

- Scale of reduction across four categories reflects common paths of grammaticalization from more content-ful to more functional elements (e.g., Hopper & Traugott 1993; Liu et al. 2010).

Conclusion

- Open questions:
  - How do these results interact with other (e.g., syntactic, semantic) criteria of lexical/grammatical contrasts? How do these criteria from different domains interact in cueing lexical/grammatical distinctions?
  - How should such non-binary divisions be represented in formal models?

Thank you!

Acknowledgements to Arto Anttila, Sharon Inkelas, Dan Jurafsky, Joan Bresnan, Rebecca Starr, and audiences at Stanford University for feedback and discussion. Special thank you to Keith Johnson for sharing his segmental distance measures for the Buckeye corpus.